New Technology Report

Voiceless Communications Technologies

Hiroyuki Manabe, Akira Hiraiwa and Toshiaki Sugimura

Talking over the mobile phone is a problem nowadays in silent environments and public places. One of the solutions to this problem is voiceless speech recognition, speech recognition in the without the voice.
This article describes the voiceless speech recognition and reports the relevant research.

1. Introduction

Nowadays, talking over the mobile phone is a problem in places where silence is required, as well as in public places. The problem caused from bad manners: talking over the mobile phone is annoying to surrounding people. The solution to this problem requires the user to talk quietly so that those around them cannot hear their voice. The challenge is that talking quietly leads to the deterioration in the signal-noise (S/N) ratio, as the noise in the background becomes relatively high. If voice can be recognized by signals other than voice signals, it should be possible to carry out communications even if the user talks so quietly that those around them cannot hear their voice. Put differently, audible voice should not be unconditionally required if speech recognition could be performed based on signals other than voice signals. NTT DoCoMo refers to this kind of speech recognition as "voiceless speech recognition" [1].

2. Background

In line with the diffusion of mobile phones in recent years, people are asked to refrain from using their mobile phones in places where silence is required such as libraries, conference rooms and offices, as well as public places including trains, buses and restaurants. This is because of the impact of mobile phones on the operation of electronic equipment such as pacemakers, but also to the issue of bad manners: people find it annoying when someone close by is talking aloud at places where silence is required and at public places.

The a survey conducted by Nikkei Research Inc. in November 1997 [2] is an example of an empirical study of peo-

ple who find mobile-phone use unpleasant. According to this survey, those who found long conversations over mobile phones in trains and buses either "extremely unpleasant" or "somewhat unpleasant" accounted for a 90% of total. Furthermore, according to a 1996 survey conducted by the Postal Services Agency of the Ministry of Public Management, Home Affairs, Posts and Telecommunications (the former Ministry of Posts and Telecommunications) [3], about 80% of the respondents have found ring tones and conversations over mobile phones unpleasant and/or annoying in the past. The same survey conducted in 2000 [4] reveals that approximately 70% of the respondents found mobile phone use annoying specifically "on the trains." These survey results show that many people find mobile phone use annoying at places where silence is required, also at public places.

The purpose of this study is to work on methods to solve this problem.

3. Voiceless Communications

The aforementioned problem may be solved by using inaudible communications technologies. This chapter describes the conventional studies on voiceless communications technologies, the profile of voiceless speech recognition currently under study, and electromyography used in this study.

3.1 Techniques based on Voice Signals/Non-voice Signals

In regard to the problem mentioned above, if the mobile phone user can talk quietly so that those around him/her cannot hear his/her voice, the impact on them should be minimal.

One way to achieve this is communications based on whispering. Acoustics studies on whispering have already been taking place, such as studies on sound analysis and speech recognition of whispering [5] and studies on synthesizing voice that would have been generated otherwise through the detection of whispering and its conversion into audible voice [6]. Such techniques that use whispering should be able to minimize the impact on people around the mobile phone user. However, the use of whispering makes it impossible to ignore the impact of noise in the background. This is particularly the case in mobile environments, where the level and source of noise in the background vary widely and are unstable. Therefore, it is believed that communication based solely on whispering would be incapable of accurately conveying the message.

An alternative to the whispering-based technique is a speech recognition technique using information other than voice signals, that is, non-voice signals. As the speech recognition technique based on non-voice signals is not affected by (acoustic) noise in the background, it is believed to be applicable to mobile environments. Moreover, the fact that voice signals are not used means that audible voice is not absolutely required; therefore, it can lessen the impact on people around the mobile phone user to an even greater degree than the whispering-based technique. This study addresses speech recognition based on non-voice signals, targeting voice generated in the same manner as sighs i.e., whispering extremely quietly and making vocalization movements without vibrating the vocal cords at all. The act of whispering extremely quietly and making vocalization movements without vibrating the vocal cords is referred to as "voiceless speech and voice" in the context of this study. Also, speech recognition based on non-voice signals is referred to as "voiceless speech recognition" in this study, as is speech recognition that is applicable to inaudible voice.

The aforementioned study referred to in Reference [5] conducts recognition, whereas the study in Reference [6] does not. While recognition is not necessarily required in human-to-human communications, if recognition is possible, it may be applied to text input and command input. Considering that the performance speech recognition will expand the range of applications, speech recognition is addressed by this study.

When a person speaks, we can confirm that the muscles around their mouth move in coordination with each other during vocalization, whether or not the voice is audible. As a result of the coordinated movement of the muscles, the person's jaw opens and closes, lips and tongue change in shape. This implies that if we can detect the movements around the person's mouth, we should be able to perform speech recognition even in the absence of audible voice. **Figure 1** shows voiceless speech recognition subject to this study.

3.2 Recognition using Electromyography

One way to detect physical movements is to use electromyography, which represents changes in electric potential in muscle cells associated with muscle activity. Electromyography is conducted by attaching two electrodes on the skin surface to detect the difference in electric potential between the two electrodes. electromyography is distinctive in that the fluctuations are violent when the muscle activity level increases but limited when it decreases (**Figure 2**). Exploiting such characteristics, studies

have been conducted for some time to estimate the muscle activity level and recognize the movements.

An example is the Cyber Finger [7], which recognizes the angles at which the fingers are bent using neural networks based on 2ch electromyography detected at the wrist. At high accuracy, it recognizes the angles at which the five fingers on one hand are bent at the ten joints. There are also studies on speech recognition based directly on electromyography, as exemplified by a study that attempts to recognize the five vowels in Japanese language by detecting 3ch electromyography on a person's face [8]. The said study extracts the electromyography every 40ms and measures the frequency of intercepting with the threshold to determine the muscle activity status in binary terms. It reports that 64% of the five vowels could be recognized on average, based on electromyography using an automaton*. There are also studies that sought to recognize ten types of English words based on 4ch electromyography and managed to recognize 60% of them on average [9]. Nonetheless, the recognition rate is insufficient in these speech recognition studies based on electromyography. Furthermore, conventional studies have not revealed whether recognition is possible in the absence of audible voice.

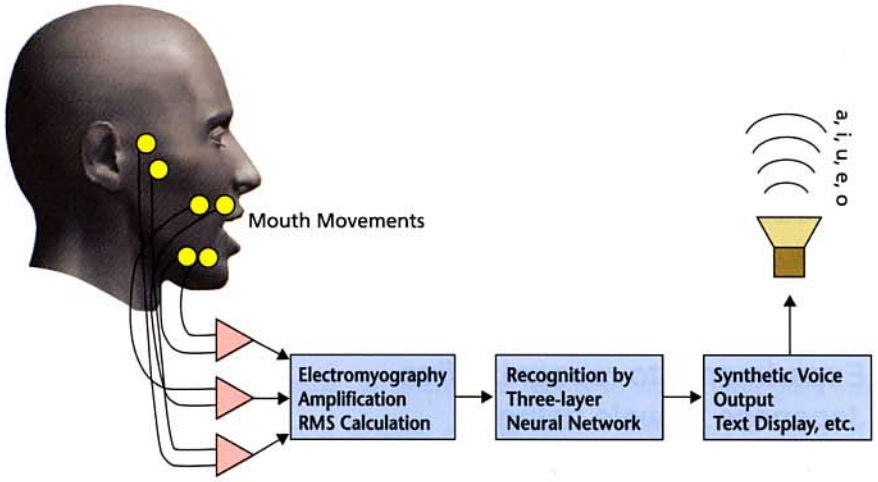

Figure 1 Voiceless Speech Recognition

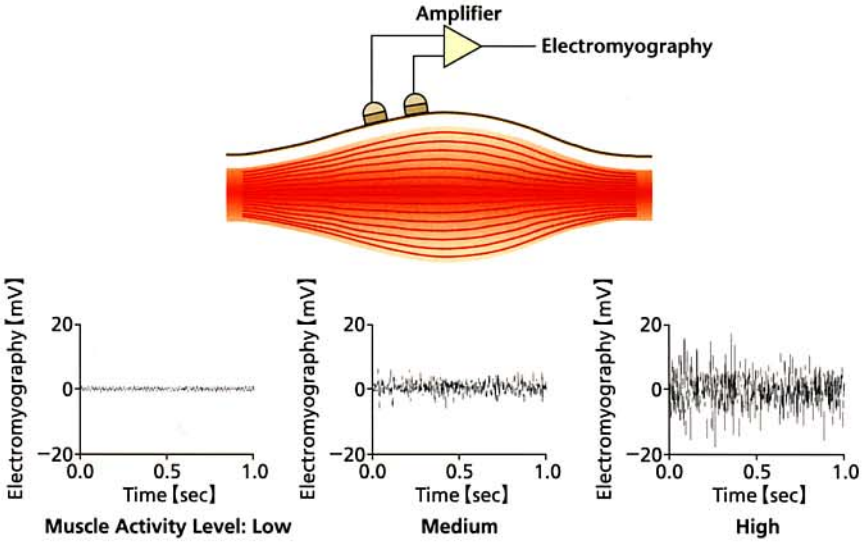

Figure 2 Measurement Method and Characteristics of Electromyography

4. Voiceless Speech Recognition using Electromyography

One way to implement voiceless speech recognition is to use electromyography. This chapter explains the reasons for adopting a method using electromyography and the approach to the study.

4.1 Affinity between Electromyography and Mobile Phones

As described earlier, voiceless speech recognition requires the detection of movements around the mouth. There are a number of ways to detect such movements, such as using picture information and magnetic information. For this study, we decided to adopt a technique using electromyography, because it is highly suitable when combined with mobile phones. Currently, a mobile phone is used by pressing it against one side of the user's face, or at least extremely close to it. As electrodes must be in contact with the skin to conduct electromyography, it might be acceptable to users if the electrodes are built inside the mobile phone. On these grounds, we decided to work on a technique using electromyography in this study.

* Automaton: A system that automatically determines the output and the subsequent internal status based on the input and the internal status.

4.2 Study Approach

There are two approaches to perform speech recognition based on electromyography: an approach based on phoneme recognition as in the study referred to in Reference [8]; and an approach based on the recognition of words as in Reference [9]. We decided to take the approach based on phoneme recognition in this study, and started with the recognition of five Japanese vowels.

5. Experiments to recognize five Japanese Vowels using Electromyography

We conducted experiments to recognize vowels using electromyography. This chapter describes the experiment method aimed at determining whether five Japanese vowels can be recognized in the absence of audible voice by using electromyography.

5.1 Muscles subject to Measurement

Firstly, it is necessary to select the muscles to be measured. To do this, we studied the muscles that are active when the five Japanese vowels are pronounced. Consequently, we determined that three muscles largely contribute to the pronunciation of the five Japanese vowels, namely, the orbiculas oris, the zygomaticus major and the digastricus. **Figure 3** shows the placement of these three muscles. They have different functions: the orbicular muscle of the mouth reduces the size, contracts and sticks out the lips; the zygomaticus major pulls back and pulls up the corners of the mouth (smiling); and the digastricus lifts the tongue bone and the root of the tongue, and fixes the tongue bone [10]. As the study referred to in Reference [8] also selects the three muscles mentioned above, it should be possible to recognize the five Japanese vowels based on these three muscles.

5.2 Extracted Characteristics and Recognition Technique

The aim of our experiments was to investigate how much could be recognized using electromyography, rather than finding what kind of characteristics and recognition techniques are suitable. Therefore, for the experiments, we adopted root mean square (RMS) as the characteristics and a back-propagation-type, three-layer neural network as the recognition technique, in consideration of the costs of calculation and the easiness of construction.

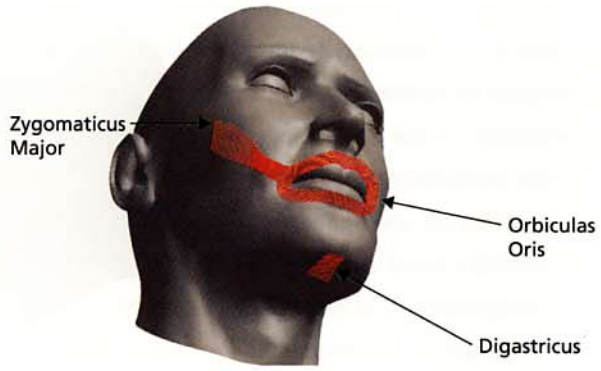

Figure 3 Muscles Placement

5.3 Electrodes

The electrodes must be in contact with the skin to conduct electromyography. There are two types of electrodes: a passive electrode, in which the electrode and the amplifier are separate to each other; and an active electrode, in which the electrode is integrated with the preamplifier. Electrodes used for medical and testing purposes are mainly passive electrodes. The drawback of passive electrodes is their tendency to absorb noise. In the experiments, we decided to use active electrodes, which, in contrast to their passive counterparts, are distinctive in that they can reduce the absorption of noise. We conducted the experiments by fixing the active electrodes with tape.

5.4 Time Window

Speech recognition using voice signals extract voice signals in time windows of tens of milliseconds and use them for analysis. Such time windows are required to recognize audible voice on a continual basis. In contrast, our experiments address the recognition of steady inaudible voice, which allows the time windows to be set wider. Specifically, we decided to extract electromyography in time windows of 400ms.

6. Recognition Experiments Results

The results of the recognition experiments mentioned above are as follows.

6.1 Non-real-time Recognition Experiment Results

Table 1 shows the results of the non-real-time (off-line) recognition experiment, performed on three subjects (all men in their 20s). Recognition was tested with respect to six states: states in which the five Japanese vowels were steadily pronounced in inaudible voice; and a steady mute state (relaxation). The five Japanese vowels pronounced steadily were represented

Table 1 Recognition Results

	relax	/aa/	/ii/	/uu/	/ee/	/oo/	average
Subject A	100	100	100	89	91	99	97
Subject B	96	92	100	97	85	73	91
Subject C	100	99	97	96	97	85	96

Unit [%]

in the form of /aa/, /ii/, …, /oo/. The experiment results indicate that the average recognition was highly accurate (more than 90%) for the three subjects.

6.2 Real-time Recognition Experiment

The experiment mentioned above was converted so that recognition could be performed at real-time. The results of this experiment also confirmed that real-time recognition could be performed at high accuracy. However, real-time recognition resulted in some mistakes upon the successive pronunciation of different vowels (for example, the pronunciation of /aa/ followed by the pronunciation of /ii/). This is believed to be attributable to the excessively large time windows.

6.3 Experiment Results

These experiments confirmed that the five Japanese vowels pronounced steadily voiceless could be recognized at high accuracy using electromyography.

7. Future Issues

In these experiments, we were able to recognize the five Japanese vowels pronounced steadily voiceless at high accuracy based on electromyography. However, the recognition was limited to vowels pronounced steadily. There are many issues to be solved in the future, as stated below.

7.1 Dynamic Recognition

In order to perform speech recognition based on electromyography, it is important to detect and recognize the temporal changes in electromyography. In the experiment, for example, mistakes were made when two vowels were pronounced in one time window. Electromyography must be dynamically analyzed and utilized for recognition in order to solve these mistakes.

7.2 Consonants Recognition

Another issue for the future is how to recognize consonants. As the transitional temporal changes in electromyography are distinctive in some consonants, the existing technique cannot directly be applied to them.

7.3 Measuring Device

Electromyography requires the electrodes to be in contact with the skin immediately above the target muscle. One way to keep the electrodes in contact with the skin is to fix them with tape –however, this will considerably undermine user convenience. An alternative way is to mount electrodes inside a mobile terminal such as mobile phones and to make the user press the terminal against his/her face. However, it is impossible to conduct electromyography from the three muscles just by mounting electrodes in existing mobile phones. Hence, we manufactured a measuring device shaped like a mobile terminal that can conduct electromyography from the three muscles targeted in the experiments (**Photo 1**). In the future, we plan to conduct experiments using the measuring device shown in Photo 1.

8. Conclusion

This article proposed voiceless speech recognition as a potential means to solve the problems associated with talking in quiet environments and public places. We studied and conducted experiments on voiceless speech recognition using elec-

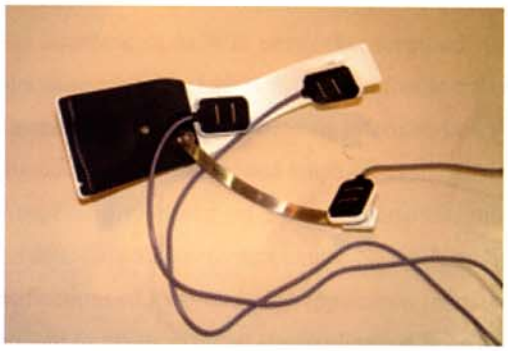

Photo 1 Measuring Device for Mobile Terminal type-Electromyography

tromyography. As a result, we revealed that the five Japanese vowels pronounced steadily voiceless could be recognized at high accuracy.

REFERENCES

[1] Manabe, Hiraiwa and Sugimura: "Speech Recognition with EMG—Vowel discrimination in steady state—," Interaction 2002, IPSJ Symposium Series, Vol.2002, No.7, pp.181-182, 2002.

[2] http://www.nikkei-r.co.jp/report/9802/tel.htm

[3] http://www.soumu.go.jp/joho_tsusin/pressrelease/japanese/denki/1015j602.html

[4] http://www.soumu.go.jp/joho_tsusin/pressrelease/japanese/denki/000413j601.html#03

[5] Ito, Takeda and Itakura: "Acoustic Analysis and Recognition of whispered speech," Technical Report of IEICE, SP2002-71, pp.59-64, 2001.

[6] http://www.ipa.go.jp/NBP/13nendo/13mito/mdata/6-23.htm

[7] Hiraiwa, Uchida, Shimohara and Sonehara: "EMG Recognition with a Neural Network Model for Cyber Finger Control," Transaction of the Society of Instrument and Control Engineers, Vo1.30, No.2, pp.216-224, 1994.

[8] N. Sugie, K. Tsunoda, "A speech prosthesis employing a speech synthesizer", IEEE Transaction on Biomedical Engineering, Vol.BME-32, No.7, pp.485-490, 1985.

[9] M.S. Morse, Y.N. Gopalan, M. Wright, "Speech recognition using myoelectric signals with neural network", Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol.13, No.4, pp. 1877-1878, 1991.

[10] John H. Warfel: "The Extremities," Lea & Febiger, 1993.